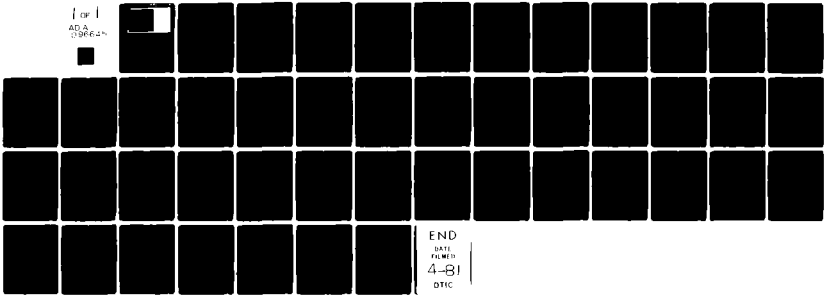
AD-A096 645 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
SAMPLING AND BAYES' INFERENCE IN SCIENTIFIC MODELING AND ROBUST--ETC(U)
DEC 80 G E BOX DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2160 NL



MRC Technical Summary Report #2160

SAMPLING AND BAYES' INFERENCE IN SCIENTIFIC MODELING AND ROBUSTNESS

George E. P. Box

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

December 1980

(Received September 4, 1980)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

SAMPLING AND BAYES' INFERENCE IN SCIENTIFIC MODELING AND ROBUSTNESS

George E. P. Box

Technical Summary Report # 2160

December 1980

ABSTRACT

Scientific learning is an iterative process employing Criticism and Estimation. Correspondingly the formulated model factors into two complementary parts -- a predictive part allowing model criticism, and a Bayes posterior part allowing estimation. Implications for significance tests, the theory of precise measurement, and for ridge estimates are considered. Predictive checking functions for transformation, serial correlation, bad values, and their relation with Bayesian options are considered. Robustness is seen from a Bayesian viewpoint and examples are given. For the bad value problem a comparison with M estimators is made.

AMS (MOS) Subject Classifications: 62A15, 62A20

Key Words: Iterative Learning; Model Building; Inference; Bayes' Theorem' Sampling Theory; Predictive Distribution; Diagnostic Checks; Transformations; Serial Correlation; Bad Values; Outliers; Robust Estimation.

Work Unit Number 4 - Statistics and Probability

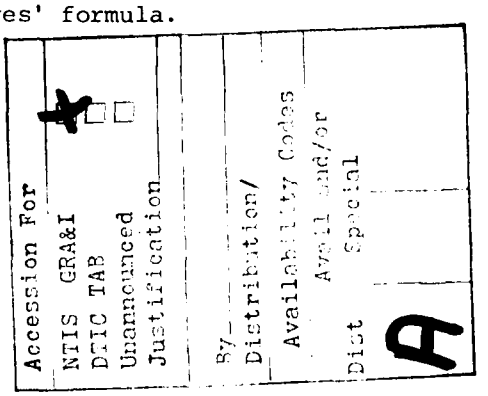
Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Scientific method is a process of guided learning in which accelerated acquisition of knowledge relevant to some question under investigation is achieved by a hierarchy of iterations in which induction and deduction are used in alternation.

This process employs a developing model (or series of models implicit or explicit) against which data can be viewed. At any given stage of the investigation, the current model approximates relevant aspects of the studied system and motivates the acquisition of further data as well as its analysis. By the use of a prior distribution it is possible to represent some aspects of such a model as if they were completely known and others as if they were more or less unknown.

Now parsimony requires that, at any given stage, the model is no more complex than is necessary to achieve a desirable degree of approximation and since each investigation is unique we cannot be sure in advance that any model we postulate will meet this goal. Therefore, at the various points in our investigation where data analysis is required, two types of inference are involved: model criticism and parameter estimation. To effect the latter, conditional on the plausibility of the model, and given the data, we can, using Bayes' Theorem, deduce posterior distributions for unknown parameters and so make inferences about them. But, before we can rely on such conditional deduction, we ought logically to check whether the model postulated accords with the data at all and, if not, consider how it should be modified. In practice, this question is usually investigated by inspecting residuals, by other informal techniques, and sometimes by making formal tests of goodness of fit. In any case this inferential procedure of model criticism whereby the need for model modification is induced, is ultimately dependent on sampling theory argument. These principles may be formalized by an appropriate analysis of Bayes' formula.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

SAMPLING AND BAYES' INFERENCE IN SCIENTIFIC MODELING AND ROBUSTNESS

George E. P. Box

No statistical model can safely be assumed adequate. Perspicacious criticism employing diagnostic checks must therefore be applied. But while such checks are always necessary, they may not be sufficient, because some discrepancies may on the one hand be potentially disastrous and on the other be not easily detectable. In addition therefore it is often pertinent to make the developing model robust against contingencies to which it is currently judged sensitive.

The object of this paper is to review the complementary roles in the model building process of the predictive distribution and of the posterior distribution; the former in producing diagnostic checks of parametric as well as residual features of the model, the latter in providing a general basis for robust estimation.

1. Scientific learning and statistical inference.

Much of statistics is concerned with extending knowledge by building empirico-mechanistic models that involve probability. A theory about such scientific model building ought to explain what good statisticians and scientists actually do. It seems that scientific knowledge advances by a practice-theory iteration. Known facts (data) suggest a tentative theory or model, implicit or explicit, which in turn suggests a particular examination and analysis of data and/or the need to acquire further data; analysis may then suggest a modified model that may require further practical illumination and so on. I shall suppose that data are acquired from a designed experiment, but the same argument would apply if data acquisition was from a sample survey or even from a visit to the library. New knowledge thus evolves by an interplay between dual processes of induction and deduction in which the model is not fixed but is continually developing. The statistician's role is to assist this evolution (see for example Box and Youle (1955), Box (1976)). In doing so he employs two inferential devices: Criticism* and Estimation.

*The apt naming of inferential criticism is due to Cuthbert Daniel. See also Popper (1959).

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

Suppose that at some stage i of an investigation, model M_i is being entertained.

Criticism can induce model modification. It involves a confrontation of M_i with available data $\underset{\sim}{y}$ (old as well as newly acquired), and asks whether M_i is consonant with $\underset{\sim}{y}$ and, if not, how not. It employs a process of diagnostic checking (see for example Box and Jenkins (1970)), which is often done informally using plots of various kinds of residual quantities, or more formally, with tests of goodness of fit or "tentative overfitting" procedures. When a modification to M_{i+1} has been made, this new model, in addition to confronting the same data, will in some cases be checked against new data generated by a design D_{j+1}. This new design will be chosen to explore those shadowy regions whose illumination is judged currently to be important in view of the nature of the modified model and the needs of independent verification.

Estimation. When the iteration leads to a model worthy to be entertained it may be used to estimate parameters conditional on its truth. In practice such estimation is used not only at the termination of the model building sequence but at many stages throughout it. This is because, to conduct criticism of a model, it is often necessary to provisionally estimate parameters at intermediate stages.

In any such iteration many subjective choices are made, conscious or unconscious, good or bad. They determine for instance which plots, displays and checks of data and residuals are looked at; and what treatments and variables are included at which levels, over what experimental region, in which transformation, in what design, to illuminate which models. The wisdom of these choices over successive stages of development is the major determinant of how fast the iteration will converge or of whether it converges at all, and distinguishes good scientists and statisticians from bad. It is in this context that theories of inference need be considered. While it is comforting to remember that a good scientific iteration

is likely to share the the property of a good numerical iteration - that mistakes often are self-correcting, this also implies that the investigator must worry particularly about mistakes which are likely not to be self-correcting.

1.1 Rival theories of inference.

The distinction between model criticism and parameter estimation has not always been made and proponents both of sampling inference and Bayesian inference have long sought for a single comprehensive theory.

I believe that, subject to some overlap discussed later, sampling theory is needed for exploration and ultimate criticism of an entertained model in the light of current data, while Bayes theory is needed for estimation of parameters conditional on the adequacy of the entertained model. On this view (see also Box and Tiao (1973)) both processes would have essential roles in the continuing scientific iteration just as the two sexes are required for human reproduction. Attempts to choose between two entities which were not alternative but complementary could certainly be expected to lead to contention, paradox, and confusion of the kind we have been experiencing. The view that more than one mode of statistical reasoning can be useful is not, of course, new and was advanced (however with a different emphasis and conclusions) by R. A. Fisher. See also in particular Dempster (1971).

1.2 The need for prior distributions.

In the past the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief. The interconnection between model assumptions and prior distributions becomes clear when it is remembered that every model can be imagined as embedded in a more complex one. For example an outright assumption of normality can be modelled by a suitable parametric family of distributions indexed by a parameter β, which has a sharp prior at the normal

value. I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters. The model is the prior in the wide sense that it is a probability statement of all the assumptions currently to be tentatively entertained a priori. On this view, traditional sampling theory was of course not free from assumptions of prior knowledge. Instead it was as if only two states of mind had been allowed--complete certainty or complete uncertainty.

One illustration of how implied prior knowledge which is implausibly imprecise can lead to trouble in sampling theory is the famous discovery by Stein (1956) of the inadmissibility of the multivariate sample mean. Consider for example the usual one-way analysis of variance set-up. The prior assumption which justifies the shrinkage estimator (see, for example, Box and Tiao (1968a), Lindley and Smith (1972)) that the group means μ_j are random samples from some normal superpopulation having unknown mean and variance might, in appropriate circumstances, be eminently reasonable. It is easy, however, to miss the lesson which is to be learned from such examples. Notice that there are many circumstances in which this "Model II" assumption would not be sensible either. For example, if the μ's were daily batch yields from some production process, it might be much more reasonable to postulate a priori that they followed some time series model such as a stationary autoregressive process. The estimators (Tiao and Ali (1971)) then derived from Bayesian means are not Stein's shrinkage estimators, but alternative estimators allowing incorporation of relevant sample information about the autocorrelation of the batch means. Thus while for this example, except as a numerical approximation, we ought not to use the sample means as estimates, we ought not to use Stein's shrinkage estimates either. There seems no logical way to avoid trouble except by the explicit prior statement of the model we wish to entertain.

1.3 Two complementary factors from Bayes' formula.

If the prior probability distribution of parameters is accepted as essential, then, a complete statement of the entertained model at any stage of an investigation is provided by the joint density for potential data $\underset{\sim}{y}$ and parameters $\underset{\sim}{\theta}$ calculated from

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A) = p(\underset{\sim}{y}|\underset{\sim}{\theta},A)\, p(\underset{\sim}{\theta}|A) \ . \tag{1.1}$$

In these expressions A is understood to indicate conditionality on all or some of the assumptions in the model specification. This model (1.1) means to me that current belief about the outcome of contemplated data acquisition would be calibrated with adequate approximation by a physical simulation involving random sampling from the distributions $p(\underset{\sim}{y}|\underset{\sim}{\theta},A)$ and $p(\underset{\sim}{\theta}|A)$.

The model can also be factored as

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A) = p(\underset{\sim}{\theta}|\underset{\sim}{y},A)p(\underset{\sim}{y}|A) \ . \tag{1.2}$$

In particular the second factor on the right, which can be computed before any data become available,

$$p(\underset{\sim}{y}|A) = \int p(\underset{\sim}{y}|\underset{\sim}{\theta},A)p(\underset{\sim}{\theta}|A)d\underset{\sim}{\theta} \tag{1.3}$$

is the predictive distribution. It is the distribution of the totality of all possible samples $\underset{\sim}{y}$ that could occur if the assumptions were true.

When an actual data vector $\underset{\sim}{y}_d$ becomes available

$$p(\underset{\sim}{y}_d,\underset{\sim}{\theta}|A) = p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A)p(\underset{\sim}{y}_d|A) \tag{1.4}$$

and the first factor on the right is Bayes' posterior distribution of $\underset{\sim}{\theta}$ given $\underset{\sim}{y}_d$

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A) \propto p(\underset{\sim}{y}_d|\underset{\sim}{\theta},A)p(\underset{\sim}{\theta}|A) \ . \tag{1.5}$$

But of equal importance is the second factor

$$p(\underset{\sim}{y}_d|A) = \int p(\underset{\sim}{y}_d|\underset{\sim}{\theta},A)p(\underset{\sim}{\theta}|A)d\underset{\sim}{\theta} \ , \tag{1.6}$$

the predictive density associated with the particular data $\underset{\sim}{y}_d$ actually obtained. Figure 1 illustrates for a single parameter θ and a sample $\underset{\sim}{y}_d$ of $n = 2$ observations.

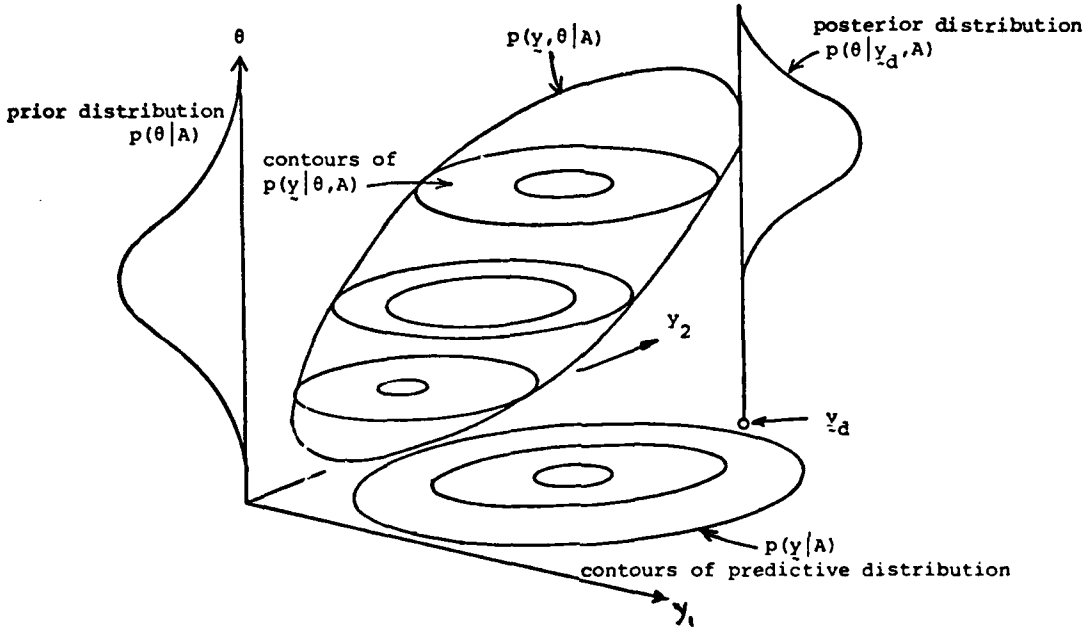


Fig. 1. A representation of the prior distribution, posterior distribution and predictive distribution, for a single parameter θ and sample y_d of two observations.

If the model is to be believed, the posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A)$ allows all relevant estimation inferences to be made about $\underset{\sim}{\theta}$. However, if $\underset{\sim}{y}_d$ were such as would be very unlikely to be generated by the model, this could not be shown by any abnormality in this factor, but <u>could</u> be assessed by reference of the density $p(\underset{\sim}{y}_d|A)$ to the predictive reference distribution $p(\underset{\sim}{y}|A)$ or of the density $p\{g_i(\underset{\sim}{y}_d)|A\}$ of some relevant checking function $g_i(\underset{\sim}{y}_d)$ to its predictive distribution. The importance of the predictive distribution and the possibility of using it in some way as a model checking device has been discussed by a number of authors. See in particular Roberts (1965), Guttman (1967), Geisser (1971, 75), Geisser and Eddy (1979), Dempster (1971, 75) and Kadane et al (1979). Also measures of surprise other than that discussed here have been proposed, for example by Good (1956).

2. <u>Estimation of the mean of a normal distribution.</u>

As an example consider a sample of n observations drawn randomly from a normal distribution with unknown mean θ and known variance σ^2 with uncertainty about the mean expressed by supposing that, a priori, θ is distributed normally about θ_0 with known variance σ_θ^2. Then <u>conditional</u> on the adequacy of the model, θ is estimated by combining data and prior information in the normal posterior distribution

$$p(\theta|\underset{\sim}{y},A) \propto (I_{\bar{y}} + I_\theta)^{\frac{1}{2}} \exp\{-\tfrac{1}{2}(I_{\bar{y}} + I_\theta)(\theta - \bar{\theta})^2\} \qquad (2.1)$$

where $I_{\bar{y}} = n\sigma^{-2}$, $I_\theta = \sigma_\theta^{-2}$ and $\bar{\theta} = w\bar{y} + (1 - w)\theta_0$ is an appropriately weighted average of $\bar{y}$ and θ_0, with $w = I_{\bar{y}}/(I_{\bar{y}} + I_\theta)$ the proportion of information coming from the data.

The predictive distribution allowing criticism of the model by contrasting data and prior information is

$$p(\underset{\sim}{y}|A) \propto \sigma^{-(n-1)}\left(\frac{\sigma^2}{n} + \sigma_\theta^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{y} - \theta_0)^2}{\frac{\sigma^2}{n} + \sigma_\theta^2}\right]\right\}. \qquad (2.2)$$

An overall predictive check is supplied by calculating

$$\alpha = \Pr\{p(\underset{\sim}{y}|A) < p(\underset{\sim}{y}_d|A)\} = \Pr\{\chi^2_n > g(\underset{\sim}{y}_d)\} \tag{2.3}$$

where

$$g(\underset{\sim}{y}_d) = \frac{(\bar{y}_d - \theta_0)^2}{\frac{\sigma^2}{n} + \sigma^2_\theta} + \frac{(n-1)s^2_d}{\sigma^2} . \tag{2.4}$$

As an example suppose the sample consists of $n = 4$ analytical tests of yield $\underset{\sim}{y}'_d = (77,74,75,78)$ performed on a single batch from an industrial process for which it is believed that the testing variance $\sigma^2 = 1$, the process mean $\theta_0 = 70$ and the batch to batch variance is $\sigma^2_\theta = 2$. We wish to estimate the mean θ for this particular batch.

In this example $\bar{y}_d = 76$, $\theta_0 = 70$, $s^2_d = 3.33$, $I_{\bar{y}} = 4$, $I_\theta = 0.5$, $w = .89$; so that, given the appropriateness of the model previously discussed, θ is estimated by the normal distribution $N(\bar{\theta},\bar{\sigma}^2)$ with $\bar{\theta} = (.89 \times 76) + (.11 \times 70) = 75.3$, $\bar{\sigma}^2 = (4 + 0.5)^{-1} = 0.22$.

However, from the predictive check

$$g(\underset{\sim}{y}_d) = \frac{(76 - 70)^2}{2.25} + \frac{3 \times 3.3}{1} = 26 \tag{2.5}$$

and

$$\alpha = \Pr\{\chi^2_4 > 26\} < .001 \tag{2.6}$$

Thus for this example the model, and hence the estimate of θ supplied by the posterior distribution $N(75.3,0.22)$, is discredited by the predictive check.

Notice the following: (a) While the posterior distribution combines information from data and prior in a manner which is entirely appropriate if the model is to be believed, the predictive distribution contrasts these two sources of information and checks their compatibility.

(b) The predictive check formalizes questions that any competent statistician would raise having been presented with the supposed form of the model and the data. The components of $g(\underset{\sim}{y}_d)$, $\{(n-1)s_d^2\}/\sigma^2$ and $(\bar{y}_d - \theta_0)^2/\{\frac{\sigma^2}{n} + \sigma_\theta^2\}$ are the standard sampling theory checking functions for contrasting an estimate of variance with a prior value and contrasting two estimates of the same mean.

(c) In making this predictive check it was not necessary to be specific about an alternative model. This issue is of some importance for it seems a matter of ordinary human experience that an appreciation that a situation is unusual does not necessarily depend on the immediate availability of an alternative.

(d) Whereas in estimating θ assuming the model to be true the posterior distribution makes use only of the single data vector $\underset{\sim}{y}_d$ that has actually occurred, by contrast, an assessment of whether the sample $\underset{\sim}{y}_d$ is likely to have occurred at all is necessarily achieved by relating $\underset{\sim}{y}_d$ to a relevant reference set of <u>all</u> data vectors $\underset{\sim}{y}$ which could have occurred with the model true.

Inspection of the global function $g(\underset{\sim}{y}_d)$ alone would rarely ensure adequate checking of the model. In this example, for instance, it would be natural to consider the individual contributions from $\bar{y}_d$ and s_d^2 not only so that they could be separately considered, but also because unusually small values of $(n-1)s_d^2/\sigma^2$ as well as unusually large ones could point to model inadequacy. Also if n were larger, we might wish to consider other functions $g_i(\underset{\sim}{y}_d)$ of the data such as moment coefficients and serial correlation coefficients which could reveal model inadequacies believed important in the current experimental situation. This could be done by referring $p\{g_i(\underset{\sim}{y}_d)|A\}$ to the predictive distribution $p\{g_i(\underset{\sim}{y})|A\}$ derived by appropriate integration of $p(\underset{\sim}{y}|A)$. Associated with these more specific checks are (possibly vague) model alternatives, the logical consequences of which are discussed in Section 4.6.

In practice, criticism of the model is often conducted by visual inspection of residual displays and other more sophisticated plots. But such a process, although

it is informal, seems to me to fall within the logical framework described above. the plots are designed to make manifest certain "features" in the data that would rarely be extreme, if the model were true. If such a feature can be described by a function $g_i(\underset{\sim}{y}_d)$, its unusualness, if formalized, would be measured appropriately by reference to $p\{g_i(\underset{\sim}{y})|A\}$.

For the above example obvious functions for checking individual features of the model are $\bar{y}$, s^2 and suitably chosen functions of standardized residuals $\underset{\sim}{r} = (r_1,\ldots,r_n)'$ with $r_i = (y_i - \bar{y})/s$, $i = 1,\ldots,n$. These would usually include the individual residuals themselves plus other functions which, depending on the context, might include checks for needed transformation, heteroscedasticity, serial correlation, "bad values", skewness and kurtosis. See for example Anscombe (1961), Anscombe and Tukey (1963), Andrews (1971a and b).

The standardized residuals can be expressed more conveniently in terms of $n - 2$ independently distributed functions obtained by making an orthogonal transformation from $\underset{\sim}{y}$ to $\underset{\sim}{Y} = (Y_1,Y_2,\ldots,Y_n)'$ with $Y_n = \sqrt{n}\bar{y}$ and then transforming to $\bar{y},s^2$ and $\underset{\sim}{u}$ where $\underset{\sim}{u}$ is a vector of $n - 2$ residual quantities $\underset{\sim}{u} = (u_1,u_2,\ldots,u_{n-2})'$ such that

$$u_j = Y_{j+1}\Big/\left\{\sum_{i=1}^{j} Y_i^2/j\right\}^{\frac{1}{2}} . \tag{2.7}$$

The Jacobian of the transformation from $\underset{\sim}{y}$ to $\bar{y},s^2,\underset{\sim}{u}$ is proportional to $(s^2)^{\frac{n-1}{2}-1}\prod_{j=1}^{n-2}\{1 + u_j^2/j\}^{-\frac{1}{2}(j+1)}$. After transformation the predictive distribution contains n elements distributed independently

$$p(\bar{y},s^2,\underset{\sim}{u}|A) = p(\bar{y}|A)p(s^2|A)p(\underset{\sim}{u}|A) \tag{2.8}$$

where

$$p(\bar{y}|A) \propto (\sigma_\theta^2 + \sigma^2/n)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\bar{y} - \theta_0)^2/(\sigma_\theta^2 + \sigma^2/n)\} \tag{2.9}$$

$$p(s^2|A) \propto (\sigma^2)^{-\frac{1}{2}(n-1)} (s^2)^{\frac{n-1}{2}-1} \exp\left\{-\tfrac{1}{2}(n-1)s^2/\sigma^2\right\} \tag{2.10}$$

$$p(\underset{\sim}{u}|A) \propto \prod_{j=1}^{n-2} \left\{1 + \frac{u_j^2}{j}\right\}^{-\frac{1}{2}(j+1)} \tag{2.11}$$

The unusualness of $g_1 = \bar{y}$, $g_2 = s^2$ and of any residual functions of interest $g_3, g_4, \ldots, g_q$ can then be assessed by computing

$$\Pr\{p(g_j|A) < p(g_{jd}|A)\} \qquad j = 1,2,\ldots,q \tag{2.12}$$

which for unimodal distributions will be tail area probabilities. For this example these would be obtained by referring

(i) $(\bar{y}_d - \theta_0)/(\sigma_\theta^2 + \sigma^2/n)^{\frac{1}{2}}$ to the Normal table

(ii) $(n-1)s_d^2/\sigma^2$ to the χ^2 table

(iii) $g_{3d}, \ldots, g_{qd}$ to reference distributions obtained by appropriate integration of $p(\underset{\sim}{u}|A)$.

These probabilities are of course affected by transformation. Thus the answer will be a little different depending for example on whether we ask a question about s or about s^2. I do not find this particularly disturbing. Slightly different questions can be expected to have slightly different answers. We now illustrate some implications.

2.1 Significance tests

Suppose σ_θ^2 is assumed small compared with σ^2/n, so that w, the relative amount of information supplied by the data, is close to zero. Then, if this model can be relied upon, the posterior distribution will be essentially the same as the prior, sharply centered at θ_0. A practical context is one where the statistician is told that the process mean is known to be θ_0 and the batch to batch variance σ_θ^2 is negligible compared with testing variance σ^2. If he believed this model, then any data $\underset{\sim}{y}$ could do very little to change his belief that $\theta = \theta_0$. However, it could

deny the relevance of this model. In particular $q_1(\underline{y}_d)$ now involves essentially the reference of $(\bar{y}_d - \theta_0)/(\sigma/\sqrt{n})$ to normal tables; the failure of this check means that the model is discredited and therefore the Bayes calculation that leads to a sharp posterior distribution at θ_0 may not logically be undertaken.

The above most satisfactorily explains to me the rationale of a significance test.

(a) The tentative model (null hypothesis) implies that θ is close to θ_0.

(b) A check on the compatibility of this model and the data, so far as the mean is concerned, is provided by reference of $(\bar{y}_d - \theta_0)/(\sigma/\sqrt{n})$ to the Normal Table.

(c) If the tail area probability is not small we do not question the model. The application of Bayes' theorem then produces a posterior distribution which is sharply centered at θ_0. We have "no reason to question the null hypothesis".

(d) If the tail area probability is small we conclude that the model which postulated that $\theta \doteq \theta_0$ is discredited by the data, i.e., the "null hypothesis is discredited".

(e) Notice too that although the failure of this check would most immediately proscribe the use of Bayes' theorem, the failure of other checks (and of that based on s^2 in particular) would also suggest the need for model modification before proceeding further.

A difficulty that this removes for me is that, as usually formulated, significance tests had seemed to provide no basis for belief. On this formulation however the significance test provides a means of discrediting a model which if accepted would inevitably imply acceptance of the belief that θ lay close to θ_0. It is admitted that this formulation does not cover all possible circumstances in which significance tests have been used (see in particular Cox (1977)), but it is arguable that other applications are best dealt with in other ways.

2.2 Precise measurement and improper priors.

Suppose now that σ_θ^2 is assumed large compared with σ^2/n, so that $1 - w$, which measures the proportion of the information about θ coming from the prior, is close to zero. Then σ_θ^2 dominates the denominator in the predictive checking function $(\sigma_\theta^2 + \sigma^2/n)^{-\frac{1}{2}}(\bar{y} - \theta_0)$ implying that the model would not be called into question by sets of data having widely different sample averages. This is the situation where we can invoke what L. J. Savage called the "theory of precise measurement" to justify the very useful numerical approximation of the posterior distribution by $N(\bar{y},\sigma^2/n)$. Now since the predictive distribution for $\bar{y}$ does not exist at the limit $1 - w = 0$ when this limiting posterior distribution is obtained, it might be argued that, when precise measurement theory is appropriate, we have a license to apply Bayes' theorem without any restraining checks on the model. Obviously however in any imaginable experimental situation there would be values of $\bar{y}$ which would rightly be regarded as implausible given the investigator's current beliefs. Thus what is really being verified is that a non-informative prior must, to make practical sense, always be proper, even though the appropriate posterior distribution can, in suitable circumstances, be numerically approximated by substituting an improper prior.

3. The normal linear model

Suppose

$$\underset{\sim}{y} \sim N(\underset{\sim}{1}\mu + \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\theta}, \underset{\sim}{\underset{\sim}{I}}_n\sigma^2) \tag{3.1}$$

with $\underset{\sim}{1}$ a vector of unities and $\underset{\sim}{\underset{\sim}{X}}$ of full rank k such that $\underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{1} = \underset{\sim}{0}$ and suppose that prior densities are locally approximated by

$$\mu \sim N(\mu_0, c^{-1}\sigma^2), \underset{\sim}{\theta} \sim N(\underset{\sim}{\theta}_0, \underset{\sim}{\underset{\sim}{\Gamma}}^{-1}\sigma^2), \{\sigma^2/\nu_0 s_0^2\} \sim \chi^{-2}(\nu_0) \tag{3.2}$$

with μ and $\underset{\sim}{\theta}$ independent but conditional on σ^2, and $\chi^{-2}(\nu_0)$ the inverted chi square distribution.

Given a sample $\underset{\sim}{y}_d$, special interest attaches to $\underset{\sim}{\theta}$ and σ^2 which given the assumptions are estimated by $p(\underset{\sim}{\theta},\sigma^2|\underset{\sim}{y}_d,A)$ with marginal distributions

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A) \propto \left\{1 + \frac{(\underset{\sim}{\theta} - \bar{\underset{\sim}{\theta}}_d)'(\underset{\sim}{X}'\underset{\sim}{X} + \underset{\sim}{\Gamma})(\underset{\sim}{\theta} - \bar{\underset{\sim}{\theta}}_d)}{(n + \nu_0)\hat{\sigma}_d^2}\right\}^{-\frac{1}{2}(n+\nu_0+k)} \tag{3.3}$$

$$p(\sigma^2|\underset{\sim}{y}_d,A) \propto \sigma^{-(n+\nu_0+2)}\exp\{-\tfrac{1}{2}(n + \nu_0)\hat{\sigma}_d^2/\sigma^2\} \tag{3.4}$$

with

$$\bar{\underset{\sim}{\theta}}_d = (\underset{\sim}{X}'\underset{\sim}{X} + \underset{\sim}{\Gamma})^{-1}(\underset{\sim}{X}'\underset{\sim}{X}\hat{\underset{\sim}{\theta}}_d + \underset{\sim}{\Gamma}\underset{\sim}{\theta}_0),\ \hat{\underset{\sim}{\theta}}_d = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y}_d,\ \nu = n - k - 1\ ,$$

$$(n + \nu_0)\hat{\sigma}_d^2 = \nu s_d^2 + \nu_0 s_0^2 + (\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)'\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{\Gamma}^{-1}\}^{-1}(\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0) + (n^{-1} + c^{-1})^{-1}(\bar{y} - \mu_0)^2\ .$$

Now let s_p^2 be the pooled estimate

$$(\nu + \nu_0)^{-1}(\nu s^2 + \nu_0 s_0^2)\ . \tag{3.5}$$

Then the predictive distributions for $(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)/s_p$, s^2, and the $\nu - 1$ elements of the residual vector $\underset{\sim}{u}$, defined in an analogous manner to that previously employed in (2.7), are independent and are given by

$$p\{(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)/s_p|A\} \propto \left\{1 + \frac{(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)'\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{\Gamma}^{-1}\}^{-1}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)}{(\nu + \nu_0)s_p^2}\right\}^{-\frac{1}{2}(n+\nu_0-1)} \tag{3.6}$$

$$p(s^2/s_0^2|A) \propto F^{\frac{1}{2}\nu-1}\left\{1 + \frac{\nu}{\nu_0}F\right\}^{-\frac{1}{2}(\nu+\nu_0)},\ F = s^2/s_0^2 \tag{3.7}$$

$$p(\underset{\sim}{u}|A) \propto \prod_{j=1}^{\nu-1}\{1 + (u_j^2/j)\}^{-\frac{1}{2}(j+1)} \tag{3.8}$$

The predictive check derived from (3.6)

$$\begin{aligned}&\Pr\{p((\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)/s_p|A) < p((\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)/s_{pd}|A)\} \\ &= \Pr\left\{F_{k,\nu+\nu_0} > \frac{(\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)'\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{\Gamma}^{-1}\}^{-1}(\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)}{ks_{pd}^2}\right\}\end{aligned} \tag{3.9}$$

is the standard analysis of variance check for compatibility of two estimates $\hat{\underset{\sim}{\theta}}_d$ and $\underset{\sim}{\theta}_0$. It was earlier proposed as a check for compatibility of prior and sample information by Theil (1963). The predictive check derived from (3.7) $\Pr\{p(s^2|A) < p(s_d^2|A)\}$ yields the F test having ν and ν_0 degrees of freedom appropriate to check whether the two estimates s_d^2 and s_0^2 are compatible. Residual checks derived from (3.8) are obtainable as before.

3.1 Ridge estimates

Now suppose the $\underset{\sim}{X}$ matrix to be in correlation form and assume $\underset{\sim}{\theta}_0 = \underset{\sim}{0}$, $\underset{\sim}{\Gamma} = \underset{\sim}{I}_k\gamma_0$, $\nu_0 \to 0$ so that $s_p^2 \to s^2$. Then the estimates $\bar{\underset{\sim}{\theta}}_d$ are the ridge estimators of Hoerl and Kennard (1970) which, given the assumptions, appropriately combine information from the prior with information from the data. The predictive check (3.9) now yields

$$\alpha = \Pr\left\{F_{k,\nu} > \frac{\hat{\underset{\sim}{\theta}}_d'\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{I}\gamma_0^{-1}\}^{-1}\hat{\underset{\sim}{\theta}}_d}{ks_d^2}\right\} \qquad (3.10)$$

allowing any choice of γ_0 to be criticized.

For example, in their original analysis of the data of Gorman and Toman (1966), Hoerl and Kennard (1970) chose a value $\gamma_0 = .25$. However substitution of this value in (3.10) yields $\alpha = \Pr\{F_{10,25} > 3.59\} < 0.01$ which discredits this choice. More recently it has been pointed out (Lindley and Smith (1972), Hoerl, Kennard and Baldwin (1975) that given the model, γ can be estimated from the data. If we do this, much smaller values of γ are obtained which of course are not in conflict with the wider model. The two kinds of analysis further illustrate the overlap between predictive checking and Bayesian estimation later discussed in Section 4.6.

The Bayes approach to ridge estimators has the characteristic advantage that the somewhat arbitrary prior assumptions, which have to be made even for compatible values of γ, are uncovered for criticism (see also Draper and Van Nostrand (1977)). If $\gamma_0^{-1} \to 0$, (3.10) yields the standard ANOVA significance test which has a detailed interpretation parallel to that set out in Section 2.1.

4. Diagnostic checks

It is useful to distinguish two kinds of checks which may be called respectively Overall or Multidirectional checks and Specific or Unidirectional checks. An example of the first would be a general inspection of residuals and the second a Durbin-Watson test for first order serial correlation. This distinction is made for example by Box and Jenkins (1970) in their discussion of the general philosophy of diagnostic checking. Concerning these two kinds of checks these authors say "... although [overall checks] can point out unsuspected peculiarities ... [they] may not be particularly sensitive. Tests for specific departures ... are more sensitive, but may fail to warn of trouble other than that specifically anticipated." The two alternatives ought properly to be regarded as extremes on some scale of dependence of checking procedures on specific alternatives. For example consider the fitting of a parametric time series model. While residuals themselves should always be inspected there are a number of way-stations between this overall but insensitive check and the device called "overfitting" in which a model is tentatively elaborated in a specific direction. Thus inspection of, and application of overall tests to, the auto-correlation function and the periodogram of the residuals while still non-specific is less general than the first device and much less specific than the second.

The model checking problem is comparable to that faced by a nation which fears aerial attack that might come from any direction but with certain rather wide zones more likely than others and certain specific directions believed especially likely. How should limited radio detection devices, which are less sensitive the less they are focused, be deployed? The best solution obviously involves some combination of wide and more specific searches, and theoretically could be achieved knowing prior probabilities and expected losses. Correspondingly, the competent statistician must, in a variety of contexts, be able to make intelligent guesses not only of what discrepancies are particularly likely, but which are potentially disastrous, and to allocate his effort accordingly. In practice this is done informally and is part of what an adequate training in statistics achieves.

4.1 Checking parametric features of the model.

In the examples considered above where sufficient statistics were available parameter preferences evidenced by proper priors were directly challenged, leading without a direct statement of alternatives to appropriate checks. When a specific set of assumptions A_1 alternative to A_0 are in mind then an appropriate checking function might also be obtained from the predictive ratio

$$p(y_d|A_1)/p(y_d|A_0). \tag{4.1}$$

We shall not explore this possibility further here, except to note that this ratio is a component in the direct assessment of Bayesian odds to which we refer briefly in Section 4.6.

4.2 Checking residual features of the model.

Residual checking functions are sometimes chosen on an ad hoc basis and sometimes using specific models. I think the best course is again to employ an iteration - this time between theory and intuition. An empirical procedure that works well invites the question: What kind of model would be needed for its justification? Such a model can then be considered for use in a wider context. For instance exponential smoothing and the "three term" controller were both empirically developed techniques found to be practically effective. ARIMA time series models are generalizations of the stochastic processes that could justify these methods (Box and Jenkins (1970)). In a similar way the practical usefulness of such things as the jackknife and cross validation implies the existence of corresponding models which are worthy of further analysis.

The distinction between parametric features of the model and residual features is of course arbitrary and a matter of convenience. In practice the needs of parsimony urge us to settle for reasonably simple models and to consider possible deviations from them. Consider now therefore an interesting but by no means unique method for obtaining an appropriate function of the data for informal or formal checking for a particular kind of deviation from a current model parametrized by a discrepancy parameter β.

Suppose the predictive distribution conditional on some specific choice of β is $p(\underset{\sim}{y}|\beta)$. Then a scaleless function of the data alone, appropriate to measure discrepancies from the value β_0 taken in the current model is provided by Fisher's score function

$$g_\beta(\underset{\sim}{y}) = \left.\frac{\partial \ln p(\underset{\sim}{y}|\beta)}{d\beta}\right|_{\beta=\beta_0} . \qquad (4.2)$$

We illustrate by considering some possible discrepancies from the standard normal linear model. First consider the model when there is no discrepancy so that $\beta = \beta_0$, and using the structure of (3.1) write

$$\underset{\sim}{\theta}' = (\mu \mid \underset{\sim}{\theta}'); \quad \underset{\sim}{X} = (\underset{\sim}{1} \mid \underset{\sim}{X}) \tag{4.3}$$

For simplicity we here suppose that the distributions of $\underset{\sim}{\theta}$ and $\ln\sigma$ are locally flat a priori so that $p(\underset{\sim}{\theta},\sigma) \doteq \text{const}\sigma^{-1}$. Then $p(\underset{\sim}{y}|\beta_0)$ is locally approximated by the singular distribution

$$p(\underset{\sim}{y}|\beta_0) \doteq \text{const } s^{-\nu} \tag{4.4}$$

where $s^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \underset{\sim}{y}'\underset{\sim}{R}\underset{\sim}{y}$ and $\underset{\sim}{R} = \underset{\sim}{I} - \underset{\sim}{M}$ with $\underset{\sim}{M} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'$. If we transform to $\hat{\underset{\sim}{\theta}}$, S, and $\underset{\sim}{u}$ then the standardized residuals $\underset{\sim}{u}$ which are functions of $\nu - 1$ angles are distributed as in (3.8) and,

$$p(\hat{\underset{\sim}{\theta}},s,\underset{\sim}{u}|\beta_0) \doteq \text{const } s^{-1}p(\underset{\sim}{u}|\beta_0) \tag{4.5}$$

To see the reasonableness of this set-up notice that by invocation of the linear model the investigator in effect predicts that the sample point $\underset{\sim}{y}$ will lie somewhere close to a hyperplane h_X spanned by the columns of $\underset{\sim}{X}$. The formulation above interprets "somewhere close to" as follows. Consider a future sample $\underset{\sim}{y}$ in relation to $(\hat{\underset{\sim}{\theta}},S)$ where $\hat{\underset{\sim}{\theta}}$ are the $k + 1$ coordinates of the projection $\hat{\underset{\sim}{y}}$ of $\underset{\sim}{y}$ on h_X, and S is the perpendicular distance of $\underset{\sim}{y}$ from h_X. Equation (4.5) says that locally any value of $\hat{\underset{\sim}{\theta}}$ is equally acceptable but that the density for the distance S will fall off inversely with S.

To obtain $q_\beta(\underset{\sim}{y})$ we need to determine how $p(\underset{\sim}{y}|\beta)$ depends on the discrepancy parameter β in the neighborhood of $\beta = \beta_0$.

4.3 A check for needed power transformation.

Especially when y_{max}/y_{min} is large some transformation of the data, for example $y^{(\lambda)} = (y^\lambda - 1)/\lambda$, might permit closer representation. Following the

approximate argument of Box and Cox (1964), with λ the discrepancy parameter and $\dot{y}$ the geometric mean of the y's, and for λ close to 1,

$$p(\underset{\sim}{y}|\lambda) \propto \dot{y}^{\nu(\lambda-1)} Q_\lambda^{-\frac{1}{2}\nu} \tag{4.6}$$

where the omitted constant does not depend on $\underset{\sim}{y}$ or on λ and where $Q_\lambda = \underset{\sim}{y}^{(\lambda)'} R \underset{\sim}{y}^{(\lambda)}$,

$$g_\lambda(\underset{\sim}{y}) = \left.\frac{\partial \ell np(\underset{\sim}{y}|\lambda)}{\partial\lambda}\right|_{\lambda=1} \doteq \underset{\sim}{z}' R \underset{\sim}{y}/s^2 = s^{-1} \sum_{i=1}^{n} z_i r_i \tag{4.7}$$

where $z_i = y_i\{1 - \ell n(y_i/\dot{y})\}$, $s^2 = \underset{\sim}{y}' R \underset{\sim}{y}/\nu$, and $r_i = \dfrac{y_i - \hat{y}_i}{s}$.

The predictive check may thus be performed by regressing the residuals $y - \hat{y}$ on the residuals $z - \hat{z}$ of the constructed variable $z = y\{1 - \ell n(y/\dot{y})\}$, which accords with a proposal of Atkinson (1973). The check can be made informally by plotting one set of residuals versus the other. More formally the distribution of $g_\lambda(\underset{\sim}{y})$ is not precisely known although an approximate level can be obtained by computer simulation.

Relation to other prooosed checks

Related checks proposed by Tukey (1949) and by Andrews (1971a) correlate the original residuals with those from the constructed variables $(\hat{y} - \bar{y})^2$ and $\hat{y}\ell n\hat{y}$ respectively. Both possess the advantage of having exactly known sampling distributions.

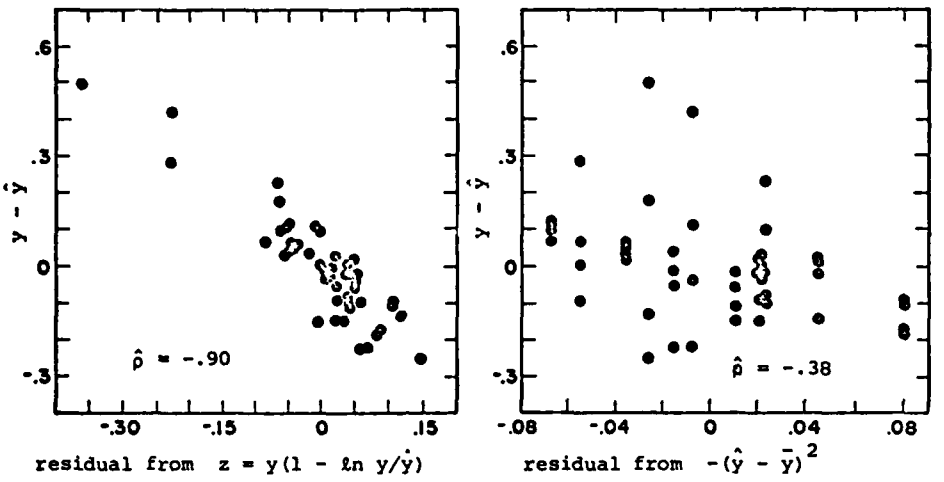
For illustration we consider

(a) the biological data of Box and Cox (1964), for which they recommend a reciprocal transformation,

(b) the trapping data of Snedecor and Cochran (1967), for which they recommend a log transformation.

Figures 2(a) and (b) show plots of residuals $y - \hat{y}$ against residuals from $y\{1 - \ell n(y/\dot{y})\}$ and $-(\hat{y} - \bar{y})^2$. The correlation coefficient for the latter transforms

(a) Box and Cox biological data



(b) Snedecor and Cochran trapping data

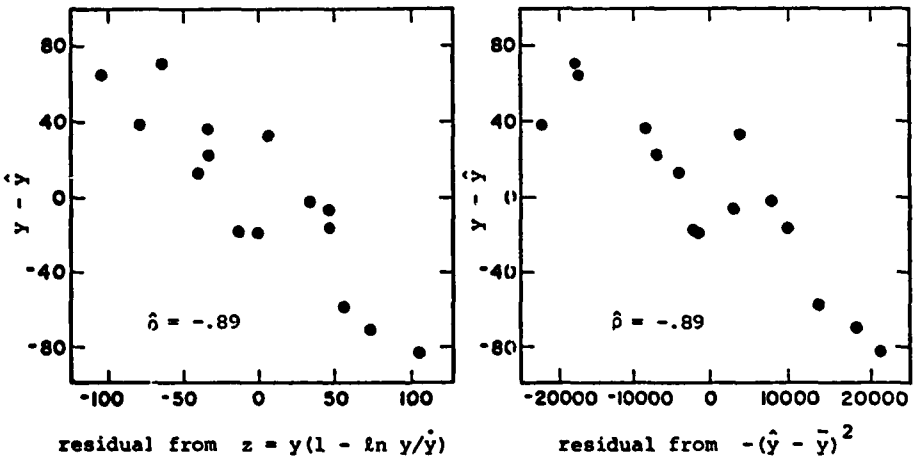


Fig. 2. Plots of residuals $y - \hat{y}$ against residuals from two constructed variables $z = y(1 - \ell n\, y/\dot{y})$ and $-(\hat{y} - \bar{y})^2$, with correlation coefficient $\hat{\rho}$ to indicate strength of association.

directly to give Tukey's one degree of freedom for non-additivity. Plots for the constructed variable $\hat{y}\ell n\hat{y}$ are not shown since they are essentially identical to the Tukey plot. The relationship between these various procedures can be seen by noting that $z = y\{1 - \ell n(y/\dot{y})\}$ may be closely approximated by

$$z \doteq \dot{y} - B(y - \dot{y})^2 \ . \tag{4.8}$$

Thus after writing

$$\frac{y_i - \dot{y}}{s} = \frac{y_i - \hat{y}_i}{s} + \frac{\hat{y}_i - \bar{y}}{s} + \frac{\bar{y} - \dot{y}}{s} = r_i + Y_i + d \tag{4.9}$$

$$g_\lambda(\underset{\sim}{y})\ \dot{\alpha} - \Sigma(r_i + Y_i + d)^2 r_i = -\{\Sigma r_i^3 + 2\Sigma r_i^2 Y_i + \Sigma r_i Y_i^2 + 2\Sigma r_i^2 d\} \tag{4.10}$$

and

$$g_\lambda(\underset{\sim}{y})\ \dot{\alpha} - (T_{30} + 2T_{21} + T_{12} + 2vd) \tag{4.11}$$

where the T_{ij} are checking functions proposed by Anscombe (1961) and Anscombe and Tukey (1963). See also Box and Cox (1964). In particular T_{12} is the component associated with Tukey's one degree of freedom for non-additivity. The approximation shows how $g_\lambda(\underset{\sim}{y})$ jointly employs skewness (T_{30}), dependence of variance on level (T_{21}), as well as transformable non-additivity (T_{12}) to indicate the need for transformation.

The Box and Cox data were generated by a 3×4 factorial with four-fold replication supplying a good deal of information about the variance as a function of location. It is not surprising therefore (see also Atkinson (1973)) that for this example $g_\lambda(\underset{\sim}{y})$ is considerably more sensitive than T_{12} (or almost equivalently, than Andrew's criterion) as a measure of the need for transformation. By contrast the Snedecor and Cochran data is from an unreplicated 3×5 arrangement where most of the information comes from T_{12} measuring non-additivity.

4.4 A check for serial correlation

For data known or suspected to have been taken in a specific serial order in time or space, a model that permitted the errors to follow a first-order autoregressive process with parameter $|\phi| < 1$ might provide an improved approximation. The dispersion matrix for the n-dimensional vector of errors $\underset{\sim}{e}$ would then be $\underset{\sim}{W}_{\phi}^{-1}\sigma^2$ where $\underset{\sim}{W}_{\phi}$ is a symmetric continuant with principal diagonal $\{1,1+\phi^2,1+\phi^2,\ldots,1+\phi^2,1\}$ and with all the elements of super- and sub-diagonals equal to $-\phi$. Thus in particular $\underset{\sim}{W}_0 = \underset{\sim}{I}$. Then

$$p(\underset{\sim}{y}|\phi) \propto (1-\phi^2)^{\frac{1}{2}} Q_{\phi}^{-\frac{1}{2}\nu} \tag{4.12}$$

where $Q_{\phi} = \underset{\sim}{y}'\underset{\sim}{R}_{\phi}\underset{\sim}{y}$ and $\underset{\sim}{R}_{\phi} = \underset{\sim}{W}_{\phi} - \underset{\sim}{M}_{\phi}$ with $\underset{\sim}{M}_{\phi} = \underset{\sim}{W}_{\phi}\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{W}_{\phi}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{W}_{\phi}$. Then with $\left.\frac{\partial \underset{\sim}{W}_{\phi}}{\partial\phi}\right|_{\phi=0} = -\underset{\sim}{C}$ where $\underset{\sim}{C}$ is $n \times n$ with unities in super and sub-diagonals and zeros elsewhere, after some algebraic manipulation

$$g_{\phi}(\underset{\sim}{y}) = \left.\frac{\partial \ell n p(\underset{\sim}{y}|\phi)}{\partial\phi}\right|_{\phi=0} \doteq \frac{1}{2}(\underset{\sim}{y}'\underset{\sim}{C}\underset{\sim}{R}\underset{\sim}{y})/s^2 \tag{4.13}$$

where $R = R_0$.

Thus

$$g_{\phi}(\underset{\sim}{y}) \doteq \sum_{i=1}^{n-1} r_i r_{i+1} \tag{4.14}$$

which is a multiple of the sample first lag autocorrelation of the residuals from the fitted model. This points to the sensitive graphical diagnostic procedure of plotting residuals r_{i+1} against r_i and yields the standard checking function of Durbin and Watson (1950).

4.5 A check for bad values

Competent investigators have over the centuries treated data as possibly containing atypical values, see for example Stigler (1973). This implies that they would not really have believed standard textbook models of the kind $y_i = f(\underset{\sim}{\theta},\underset{\sim}{x}_i) + e_i$ $(i = 1,2,\ldots,n)$ which state that the same structure is appropriate for <u>every one</u> of a sample of n observations.

When it is unknown which observations are dubious a more credible "contaminated" model proposed by Jeffreys (1932), Dixon (1953) and by Tukey (1960) supposes that there is a probability α that any given observation is "bad" (cannot be represented by the ideal model). Given α, let $p(\underset{\sim}{y}|\alpha)$ be the predictive distribution and let $p(b|\alpha)$ denote the probability of getting b bad values, then (Box and Tiao (1968b), Bailey and Box (1980a))

$$p(\underset{\sim}{y}|\alpha) = \sum_{b=0}^{n} \binom{n}{b} \alpha^{b}(1 - \alpha)^{n-b} p(\underset{\sim}{y}|b) \tag{4.15}$$

and

$$g_{\alpha}(\underset{\sim}{y}) = \left.\frac{\partial \ell n p(\underset{\sim}{y}|\alpha)}{\partial \alpha}\right|_{\alpha=0} = n\left\{\frac{p(\underset{\sim}{y}|b = 1)}{p(\underset{\sim}{y}|b = 0)} - 1\right\} . \tag{4.16}$$

Now let z_i indicate that the ith observation is bad, then

$$p(\underset{\sim}{y}|b = 1) = \frac{1}{n} \sum_{i=1}^{n} p(\underset{\sim}{y}|z_i) \tag{4.17}$$

so

$$g_{\alpha}(\underset{\sim}{y}) = \frac{\sum_{i=1}^{n} p(\underset{\sim}{y}|z_i)}{p(\underset{\sim}{y}|b = 0)} - n .$$

Depending on experimental circumstances, there are a variety of ways in which bad values might be modeled. In particular, contamination could come from increased error variance, unknown bias, and mistaken sign. The last possibility was suggested by Barnard (1978) to account for two suspiciously large outliers in Darwin's data on cross and self fertilized plants, quoted by Fisher (1935).

For illustration consider the first possibility. With one bad value, suppose the error covariance matrix $W_i^{-1}\sigma^2$ has all elements equal to σ^2, except for the i^{th} element which is equal to $\kappa^2\sigma^2 (\kappa > 1)$. Then

$$g_{\alpha}(\underset{\sim}{y}) = \frac{1}{\kappa}\left(\frac{n}{n - q}\right)^{\frac{1}{2}} \sum_{i=1}^{n} \left(\frac{s^2}{s_i^2}\right)^{-\frac{\nu}{2}} - n \tag{4.18}$$

where $$\nu s^2 = y'Ry,\ \nu s_i^2 = y'R_i y = y'\{W_i - W_i X(X'W_i X)^{-1}X'W_i\}y \tag{4.19}$$

where $W_i = I - qG_i$, $q = 1 - \frac{1}{\kappa^2}$ and G_i is an $n \times n$ matrix with a single unity for the ith diagonal element and all other elements zero. Now

$$\nu s_i^2 = \nu s^2 - \frac{q}{1 - qv_i}(y_i - \hat{y}_i)^2$$

where $v_i = \mathrm{Var}(\hat{y}_i)/\sigma^2 = x_i'(X'X)^{-1}x_i$ and $y_i - \hat{y}_i$ is the ith residual from the ideal model fit, $y - \hat{y} = Ry$.

Thus finally $g_\alpha(y) = \kappa^{-1}\left\{\frac{n}{n-q}\right\}^{\frac{1}{2}} D - n$ where

$$D = \Sigma\left\{1 - \frac{q}{\nu(1 - qv_i)}r_i^2\right\}^{-\frac{\nu}{2}} \tag{4.20}$$

This is the simplest form for computation. The nature of this checking function D can however be more clearly seen by writing it in terms of the weighted residuals $\tilde{r}_i = (y_i - \tilde{y}_i)/s_i$ where $y - \tilde{y} = R_i y$. Thus

$$D = \Sigma\left\{1 + \frac{q(1 - qv_i)}{\nu}\tilde{r}_i^2\right\}^{\frac{\nu}{2}}.$$

Thus D is proportional to the sum of the reciprocals of the n residual t ordinates obtained by downweighting (omitting as $q \to 1$) each observation in turn and recomputing the fitted value $\tilde{y}_i$ and the standard deviation s_i.

The situation of most interest is when κ is large (say $\kappa \geq 5$). Then q approaches unity and the check may be carried out by calculating

$$D = \Sigma\left(1 - \frac{r_i^2}{\nu(1 - v_i)}\right)^{\frac{\nu}{2}} = \Sigma\left\{1 + \frac{(1 - v_i)}{\nu}\tilde{r}_i^2\right\}^{\frac{\nu}{2}}. \tag{4.21}$$

Equation (4.16) brings out a feature of the checking function (4.2) which can be a disadvantage. Differentiation at $\alpha = 0$ on the boundary of the parameter space ensures that only the possibility of one bad value is taken account of. Thus as is

clear from (4.21) D in its present form would not be expected to be sensitive to the occurrence of two or more bad values. Thus with $\kappa \approx 5$, we obtain the value $D = 59.05$ for Darwin's data. A Monte Carlo study with 5000 samples of 15 observations shows that this value would be exceeded by chance about 14% of the time, which hints only mildly at inadequacy in the standard model, confirming D's insensitivity for this example.

4.6 Bayesian options for specific alternatives.

When concrete alternatives are in mind, Bayesian options are available. In particular the predictive ratio $p(\underset{\sim}{y}|A_1)/p(\underset{\sim}{y}|A_0)$, mentioned earlier, is a component in the posterior odds ratio which with suitable priors might be used to assess directly the relative evidence for one model versus another. Also $g_\beta(\underset{\sim}{y})$ of (4.2) has a Bayesian interpretation for, if corresponding to some discrepancy parameter β, the prior distribution $p(\beta)$ was locally flat then the posterior distribution $p(\beta|\underset{\sim}{y}_d)$ would be proportional to the predictive density $p(\underset{\sim}{y}_d|\beta)$ regarded as a function of β. Furthermore if that posterior distribution was approximated by a normal distribution, then

$$g_\beta(\underset{\sim}{y}) \doteq \frac{-(\beta_0 - \hat{\beta})}{\sigma_\beta^2} \tag{4.22}$$

and a second differentiation would produce a standarized variate.

The relation shows how any specific predictive check $g_\beta(\underset{\sim}{y})$ is linked to a posterior distribution. In particular, considering the illustrative examples of the last section, the marginal posterior distribution for λ was given by Box and Cox (1964), for ϕ by Zellner and Tiao (1964), for the ridge regression parameters by Lindley and Smith (1972) and that for α may be obtained using the results of Box and Tiao (1968b) and Bailey and Box (1980a).

Before leaving the topic of diagnostic checks two final points need to be made:

(i) The above discussion illustrates the "overlap" previously mentioned when specific alternatives are in mind. It does not however establish the omnipotence

of purely Bayesian inference. However far the process of model elaboration is taken by Bayesian methods the final model involving say the mth set of assumptions A_m can still be factored

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A_m) = p(\underset{\sim}{\theta}|\underset{\sim}{y},A_m)p(\underset{\sim}{y}|A_m) \tag{4.23}$$

thus there always remains an unexplored n-dimensional predictive distribution $p(\underset{\sim}{y}|A_m)$ in relation to which a small relative value for $p(\underset{\sim}{y}_d|A_m)$ could, on a sampling theory argument, discredit the assumptions on which the Bayes analysis was conditional. The same is true of the more plausible of two models chosen using a posterior odds ratio.

(ii) In addition to possible discrepancies to which we have been alerted by experience, other features may appear pointing to inadequacies of a kind not previously suspected. This possibility has sometimes proved perplexing, for while on the one hand the truly unexpected could point the way to precious new knowledge, on the other, associated probabilities would be indeterminate because of the uncountable character of other features that might also have been regarded as surprising. I think the calculation which ignores this difficulty of indeterminate selection is still worth making, for at least it helps to correct a misjudgement of something that appears unusual but really is not. For example, Feller (1968) shows that for a random group of 30 people, the probability that at least two have coincident birthdays is over 70%; this tells us we need look no further for an explanation when we are surprised to find two such people at a party. While the proposed policy will lead to the too frequent pursuit of nonexistent assignable causes, the iterative process will quickly terminate this chase.

5. Robust estimation.

Efficient iterative model building requires both diagnostic checking and model robustification, where by robustification I mean judicious and grudging elaboration of the model to ensure against particular hazards (see also Box (1979)). Robustification

becomes necessary when it is known that likely, but not easily detectable, model discrepancies can yield badly misleading analyses. It is well known, for example, that least squares analysis can be dramatically affected by moderate serial correlation of errors.

Recently the serious consequences of bad values on standard least squares analysis has been especially emphasized and numerous authors have proposed methods which rely on abandonment or modification of classical estimation methods. In discussing the rationale for this approach Huber (1977) says "The traditional approach to theoretical statistics was and is to optimize an idealized model and then to rely on a continuity principle: what is optimal at the model should be optimal nearby. Unfortunately, this reliance on continuity is unfounded: the classical optimal procedures tend to be discontinuous in the statistically meaningful topologies."

He then quotes a motivating example given by Tukey (1960), who pointed out for example that if a normal distribution were very mildly contaminated with another which is centrally located but of larger variance, then the sample standard deviation could be a very poor estimate of scale. Tukey's contribution was remarkable because it had previously gone unnoticed that the assumption that the same structure must apply to <u>every</u> observation y_i $(i = 1,2,\ldots,n)$ <u>with absolute certainty</u> $(1 - \alpha = 1)$, not only was unrealistic (since no responsible investigator would make the claim that inadvertent bad values were impossible), but also could have serious consequences. While Huber goes on to say that typical 'good data' samples in the physical sciences appear to be well modeled by this contaminated normal model, he does not develop methods based on this more realistic set up. This is presumably because his objection would apply equally to the new as well as to the old model.

I do not agree that the example would support a thesis of the need to abandon model-based procedures. A model that omits the parameter α is, of course, the same as one that includes it but sets its value exactly to zero. A value of $\alpha = 0$, which allows no possibility whatever for bad values, and a value of $\alpha = 0.001$ are, I think, not close in any statistically meaningful topology. Although 0.001 may look close to zero, an odds ratio of $0.999/0.001 = 999$ for a "good" to a "bad" value is obviously very different from one of infinity. Such differences in probability distinguish, for example, a lifeless world in which no evolution could possibly occur from the one we live in.

The proper conclusion to draw from Tukey's example is, I think, that for many practical situations in which occasional bad values are to be expected the standard linear model provides an inadequate approximation that is potentially misleading and therefore the model should be appropriately changed to approximate what is believed rather than what is not. The situation is logically the same for a model that implicitly insists there can be no serial correlation, when data have in fact been collected serially, or that no transformation of y could be needed when y_{max}/y_{min} is large. As in the classical Stein problem if we know something a priori it may be disastrous to omit it. On this view for robust estimation of the parameters of interest we should modify the model which is at fault, rather than the method of estimation which is not.

5.1 Bayesian robust estimation.

As was argued for example by Box and Tiao (1964), all relevant aspects of the problem are brought out in an appropriate Bayes analysis. Supposing that $\underset{\sim}{\theta}$ has the same physical interpretation for all β then estimation of $\underset{\sim}{\theta}$ which is robust relative to the discrepancy parameter β is supplied by the posterior distribution

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}) = \int p(\underset{\sim}{\theta}|\beta,\underset{\sim}{y})p_u(\beta|\underset{\sim}{y})p(\beta)d\beta \qquad (5.1)$$

This expression contains three key elements that repay individual study:

(a) the sensitivity of inferences about $\underset{\sim}{\theta}$ to changes in β is reflected by $p(\underset{\sim}{\theta}|\beta,\underset{\sim}{y})$ considered as a function of β;

(b) the information about β coming from the data itself is reflected in the pseudo-likelihood $p_u(\beta|\underset{\sim}{y}) = p(\beta|\underset{\sim}{y})/p(\beta) \propto p(\underset{\sim}{y}|\beta)$; (5.2)

(c) the probability of occurrence of different values of β in the real world is represented by $p(\beta)$ which can be chosen to approximate what is believed or feared.

This route was used to explore deviations from the standard normal model for a particular class of heavy-tailed distributions by Box and Tiao (1962, 64); for the contaminated model of Section 4.5 by Jeffreys (1932) and Box and Tiao (1968b); for a serial correlation model by Zellner and Tiao (1964); for a transformation problem by Box and Cox (1964). Notice that using this approach the parameters $\underset{\sim}{\theta}$ of interest are completely estimated in the sense that their distribution rather than merely a point estimate is available. Also the various elements of $p(\underset{\sim}{\theta}|\underset{\sim}{y})$ which can be studied individually can provide a deep understanding of each robustness problem. A particularly informative display shows contours of the joint distribution $p_u(\theta,\beta|\underset{\sim}{y})$ for some parameter θ of interest and the discrepancy parameter β together with the marginal distribution $p_u(\beta|\underset{\sim}{y})$. When a less prodigal display is necessary the mean and standard deviation of $p(\theta|\beta,\underset{\sim}{y})$ may be shown with $p_u(\beta|\underset{\sim}{y})$. For illustration we consider some serial data analyzed by Coen, Gomme and Kendall (1969). They regressed quarterly values of the Financial Times Share Index y on detrended lagged values of U.K. car production X_1, and of the Financial Times Commodity Index X_2 using a model* which could be written (Box and Newbold (1971)) as

*For the present purpose we retain the model structure of Coen, Gomme and Kendall. However its relevance seems dubious, for example, a multivariate time series analysis by Tiao and Box (1980) for these three series shows the stock prices y acting as a weak leading indicator for the commodity index X_2.

$$y_t = \beta_0 + \beta_1 t + \theta_1 X_{1,t-6} + \theta_2 X_{2,t-7} + e_t \quad \text{with} \quad e_t = \phi e_{t-1} + a_t \qquad (5.3)$$

with a_t white noise, and ϕ constrained to be equal to zero. Figure 3 illustrates an analysis made by Pallesen (1977) in which ϕ is unconstrained. It shows the joint posterior distribution for θ_1 and ϕ and the marginal distribution for ϕ assuming locally flat priors for $\underset{\sim}{\theta}$, $\ell n\sigma$ and ϕ. Although for this example serial correlation could have been easily detected by diagnostic checks, notice the enormous shift (about five standard deviations) of the conditional distribution $p(\theta_1|\phi,\underset{\sim}{y})$ which occurs as ϕ changes from zero to more plausible values. This illustrates the point that smaller serial correlation, of a magnitude difficult to detect with diagnostic checks, could disastrously invalidate estimates of $\underset{\sim}{\theta}$.

A second example discussed more fully in Bailey and Box (1980b) further illustrates this approach for the "bad value" problem using the contaminated normal model of Section 4.5. The data were used originally by Box and Behnken (1960) to illustrate the analysis of a balanced incomplete four factor three-level design with $n = 27$ observations arranged in three blocks of nine. A residual plot suggests the possibility of two bad values (y_{10} and y_{13}). However the small number of residual degrees of freedom and the nature of this particular design would induce large correlations yielding potentially misleading residual patterns.

Table 1 gives Bayesian means and standard deviations for coefficients in the fitted model

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \beta_{ii} x_i^2 + \sum_{i=1}^{4} \sum_{j>i}^{4} \beta_{ij} x_i x_j + e \qquad (5.4)$$

In this analysis κ was set equal to 5 and the values of α varied over the range 0 to 0.091. It has been shown by Chen and Box (1979) that for $\kappa \geq 5$ the posterior distribution is mainly a function of $\varepsilon = \alpha/(1-\alpha)\kappa$ so the results are also labelled in terms of this dominant discrepancy parameter ε. It will be noticed:

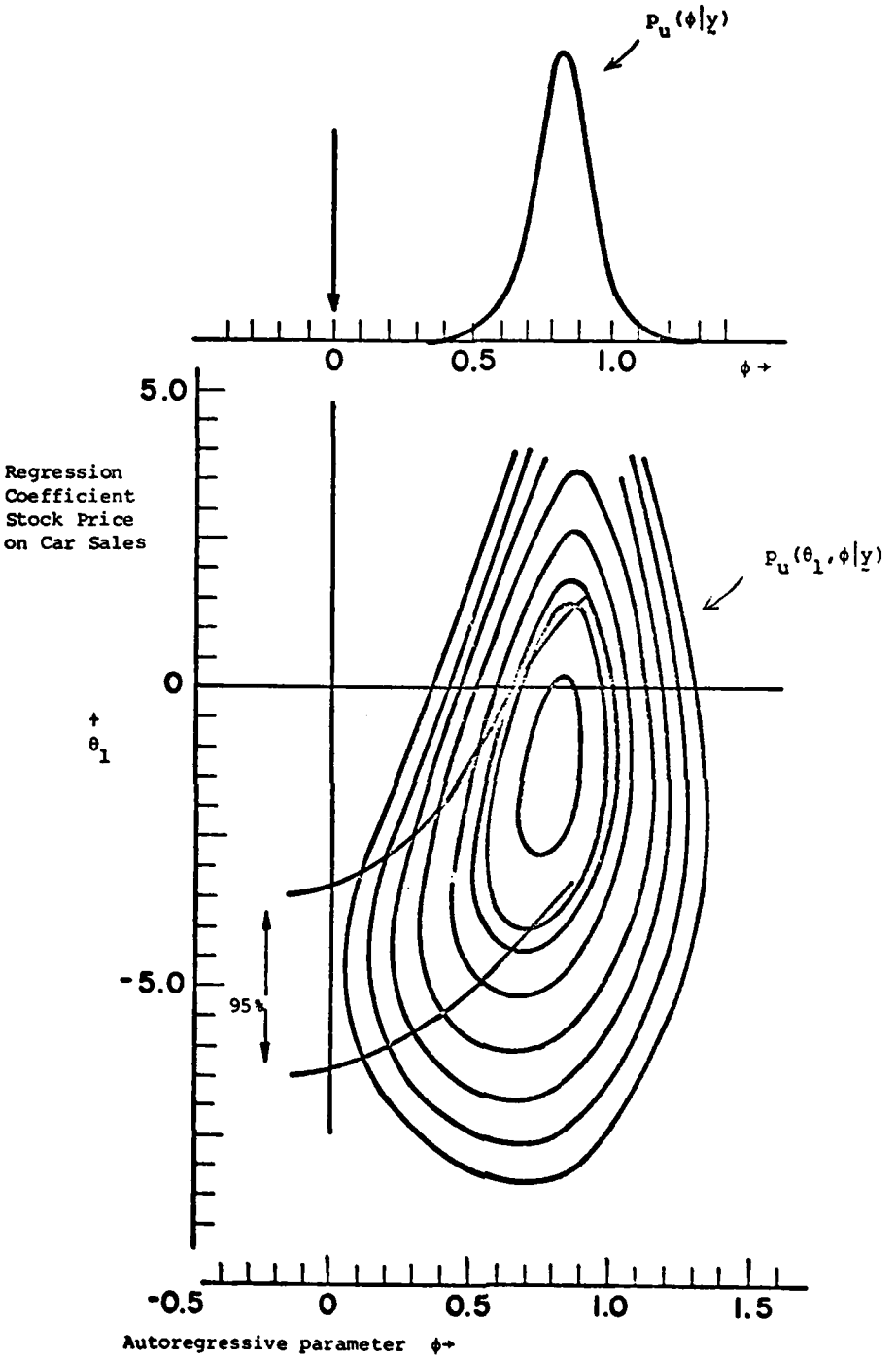


Fig. 3. Joint posterior distribution of θ_1 and ϕ and marginal posterior distribution of ϕ. Note shift in approximate 95% interval as ϕ is changed.

(a) The large change in each estimated effect and standard deviation occurs when no possibility whatever of bad values ($\epsilon = 0$) is replaced by a small possibility ($\epsilon = 0.001$). For good data the typical behavior of a table of this kind is that only very minor changes in mean and standard deviation occur as ϵ is changed over the plausible range.

(b) For all the estimates except β_{14} the standard deviations of effects are about halved. Thus for these effects the use of the more appropriate model is equivalent to a four-fold increase in the size/sensitivity of the experiment. This may be compared for example with a parallel analysis by Box and Cox of their biological data where a three-fold increase in sensitivity resulted from the use of an appropriate transformation.

(c) The analysis can be further illuminated by considering other available quantities. In particular a plot of the probability that the ith value is bad, given that one value is bad, (see for example Abraham and Box (1978)), results in a plot with 94% of the probability associated with the tenth observation and the remainder spread among the remaining 26 observations. It is likely therefore that y_{10} alone is a bad value. It is a deficiency of the design being used here that least squares estimates of interactions employ only four of the 27 observations and so lack robustness to bad observations (see for example Box and Draper (1975)). In particular $\hat{\beta}_{14} = 0.25\ (y_{10} - y_{11} - y_{12} + y_{13})$ so that the Bayesian down-weighting of y_{10} accounts for the large change in this estimate and the <u>increase</u> in the standard deviation.

(d) We saw in the case of ridge regression how failure to take account of observational information could lead to an unrealistic choice of the discrepancy parameter γ. To complete the picture therefore, a plot of the marginal distribution of the discrepancy parameter ϵ should be made in conjunction with Table 1 (compare also with the serial correlation example in Figure 3). For this data the distribution $p_u(\epsilon|\underline{y})$ has its mode close to $\epsilon = 0.010$.

	ϵ	0	.001	.005	.010	.015	.020
	α	0	.005	.024	.048	.070	.091
β_0		90.60 (.94)	90.60 (.45)	90.60 (.41)	90.60 (.41)	90.60 (.41)	90.60 (.41)
β_1		1.93 (.47)	2.46 (.28)	2.49 (.23)	2.49 (.22)	2.49 (.22)	2.49 (.22)
β_2		-1.96 (.47)	-1.96 (.22)	-1.96 (.20)	-1.96 (.20)	-1.96 (.20)	-1.96 (.20)
β_3		1.13 (.47)	1.13 (.22)	1.13 (.20)	1.13 (.20)	1.13 (.20)	1.13 (.20)
β_4		-3.68 (.47)	-3.15 (.28)	-3.12 (.23)	-3.12 (.22)	-3.12 (.22)	-3.12 (.22)
β_{11}		-1.42 (.70)	-1.88 (.44)	-1.90 (.41)	-1.90 (.41)	-1.89 (.42)	-1.89 (.42)
β_{22}		-4.33 (.70)	-4.10 (.36)	-4.09 (.34)	-4.09 (.34)	-4.09 (.34)	-4.09 (.34)
β_{33}		-2.24 (.70)	-2.01 (.38)	-2.00 (.34)	-2.00 (.34)	-2.00 (.34)	-2.00 (.34)
β_{44}		-2.58 (.70)	-3.05 (.44)	-3.06 (.41)	-3.06 (.41)	-3.06 (.42)	-3.05 (.42)
β_{12}		-1.67 (.81)	-1.67 (.39)	-1.67 (.35)	-1.67 (.35)	-1.67 (.34)	-1.67 (.34)
β_{13}		-3.83 (.81)	-3.82 (.39)	-3.82 (35)	-3.82 (.35)	-3.82 (.34)	-3.82 (.34)
β_{14}		.95 (.81)	-.45 (.95)	-.51 (.92)	-.50 (.93)	-.49 (.95)	-.48 (.95)
β_{23}		-1.67 (.81)	-1.67 (.39)	-1.67 (.35)	-1.67 (.35)	-1.67 (.35)	-1.67 (.35)
β_{24}		-2.62 (.81)	-2.62 (.39)	-2.62 (.35)	-2.62 (.35)	-2.62 (.35)	-2.62 (.35)
β_{34}		-4.25 (.81)	-4.25 (.39)	-4.25 (.35)	-4.25 (.34)	-4.25 (.34)	-4.25 (.34)

Table 1 Bayesian means and (standard deviations) for polynomial coefficients using the contaminated model of Section 4.5 with $\kappa = 5$ ($\epsilon = \alpha/(1 - \alpha)\kappa$).

The Bayes approach to robust estimation has the advantage of generality; furthermore it clearly reveals at any given stage, on precisely what assumptions the analysis is conditional. With the increased speed of computers and availability of visual display equipment a general Bayesian computer program, that can analyze any model we wish to entertain, seems a much more attractive prospect than the fresh devising of semi ad hoc procedures for each new possibility.

Some parallels in the two approaches are briefly considered below for the "bad value" problem.

5.2 Robust estimation for the "bad value" problem.

For the "bad value" problem a wide variety of semi-empirical estimators have been proposed. Among these are the M,L, and R, and various kinds of adaptive estimators. In turn among the M estimators a number of different "ψ" functions have been suggested leading to different ways of downweighting extreme observations.

Now consider the model of Section 4.5 for the simple location structure $E(y_i) = \mu$. Then (see for example Box and Tiao (1968b)) the Bayesian mean may be written

$$\hat{\mu} = \sum_{b=0}^{n} p(b|\underset{\sim}{y},\alpha)\bar{y}^{(b)} \tag{5.5}$$

where $p(b|\underset{\sim}{y},\alpha)$ is the posterior probability that there are b bad values and $\bar{y}^{(b)}$ is the corresponding conditional posterior mean. Consider in particular $\bar{y}^{(1)} = \sum w_i y_i$. Then Chen and Box (1979) show that for $\kappa \geq 5$

$$w_i \doteq (n-1)^{-1}(1 - D_i/D) \tag{5.6}$$

$$D_i = \left\{1 - \frac{nr_i^2}{(n-1)^2}\right\}^{-\frac{n-1}{2}} = \left\{1 + n^{-1}\tilde{r}_i^2\right\}^{\frac{n-1}{2}} \tag{5.7}$$

where r_i and $\tilde{r}_i$ are unweighted and weighted residuals defined in Section 4.5.

Figure 4(a) and (b) show plots of $w = w_1$ against r_1 and $\tilde{r}_1$ for three random normal samples of ten observations from a normal distribution when a multiple $0,1,2,\ldots,$ of σ is added to the first observation in each sample. Empirical approximations for these weighting curves are provided by the functions

$$w = .1 \exp\{-|.49\, r_1|^7\} \quad \text{and} \quad w = .1 \exp\{-|.3\, \tilde{r}_1|^{3.5}\}$$

Also shown in Figure 4(b) for comparison is Tukey's biweight function $w = .1\left\{1 - \left(\frac{\tilde{r}}{c}\right)^2\right\}^2$ for $c = 5.3$ (chosen to roughly match the curve). Although the Bayesian weights are sample dependent they remain remarkably stable as is indicated (a) by the smooth manner in which the remaining weight is evenly spread throughout the non-discrepant observations; (b) by the closeness with which points from different samples follow the same curve.

The estimate $\hat{\mu}$ is sample adaptive in another more striking way however. For illustration consider the case where the $p(b|\underset{\sim}{y},\alpha)$ are negligible for $b \geq 2$. Then writing $p = p(1|\underset{\sim}{y},\alpha)$ (5.5) becomes

$$\hat{\mu} = (1 - p)\bar{y} + p\bar{y}^{(1)} \tag{5.8}$$

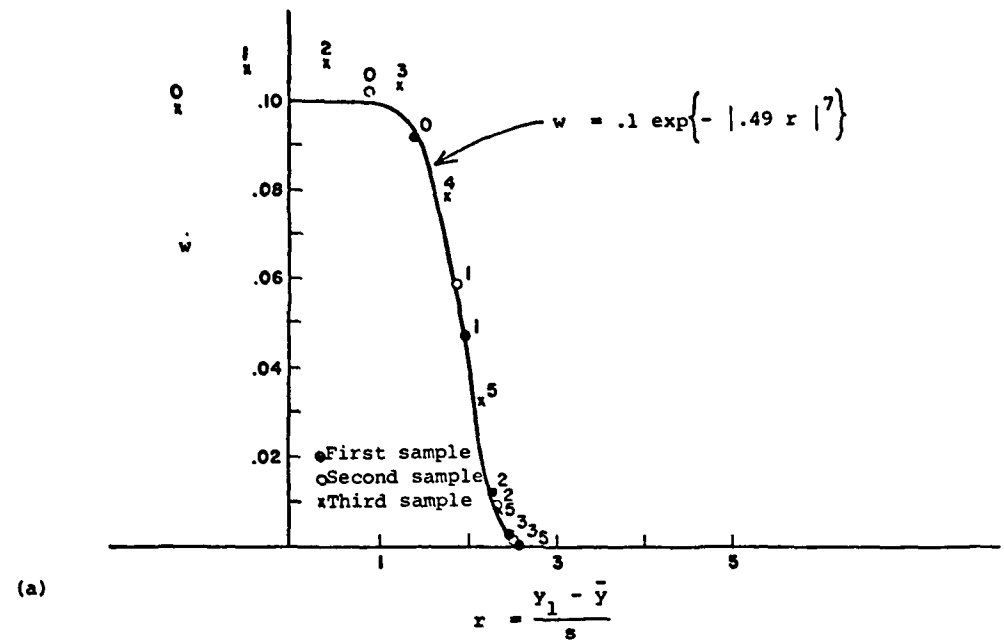
and the Bayesian mean is an interpolation between $\bar{y}$ and the "robustified" $\bar{y}^{(1)}$. In this expression the value of p is determined by the posterior odds ratio for one versus no bad values

$$p/(1 - p) \doteq \varepsilon\{n/(n - 1)\}\, D \qquad \varepsilon = \alpha/(1 - \alpha)\kappa \tag{5.9}$$

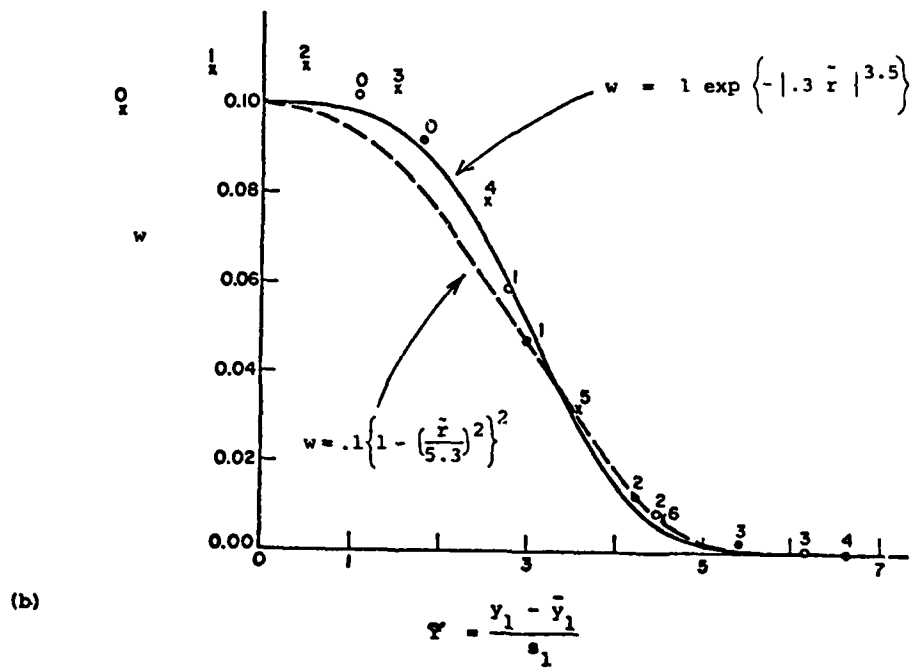
and D is the checking function encountered earlier.

Sample adaptivity is evidenced as follows. For a sample with no outliers $\bar{y}$ and $\bar{y}^{(1)}$ are not very different so that $\hat{\mu}$ is close to $\bar{y}$. But in the presence of an outlier of larger and larger size two things happen: the outlier is down-weighted in $\bar{y}^{(1)}$ which becomes more and more different from $\bar{y}$ and also p becomes larger placing more and more emphasis on $\bar{y}^{(1)}$.

The purpose of this discussion is to show that sensible solutions which appropriately downweight suspected bad values may be obtained directly from an appropriate



(a)



(b)

Fig. 4. Weight w applied to y_1 for three samples from a Normal Distribution. Numbers 0,1,2,... indicate that $0,\sigma,2\sigma,\ldots$ has been added to y_1.

model. From the viewpoint of the traditional M estimator, the weight function (W say) for $\hat{\mu}$ itself is an interpolation $W = (1 - p)\frac{1}{n} + pw$ between $1/n$ and w. Thus W will descend to the value $(1 - p)/n$ for large $\tilde{r}$. For a sample containing a large outlier, $1 - p$ will be negligible and W will approach w plotted in Figure 4(b) and will descend like Tukey's biweight. However for a more normal-looking sample W will flatten out to some moderate non-zero value and will more closely resemble the weighting originally proposed by Huber.

In choosing robust estimators there is room for empiricism but I think that some of its inspiration should be applied to the choice, study, and consequences of appropriate parsimonious models. The structure of the resulting Bayesian analysis should in each case be carefully analyzed, for the great strength of such a model-based approach is that the exact consequences of whatever goes into the model must come out. These consequences will either agree with "common sense" or they will not. If they do not then we know either that what went in was inappropriate in a way we had failed to foresee, or else, as happens quite frequently, that our common sense was too shortsighted. In either case we learn something.

ACKNOWLEDGEMENTS

This research was sponsored by the United States Army under Contract No. DAAG29-80-C-0041. My thanks are due to Steve Bailey, Gina Chen, Conrad Fung, Kevin Little, and Ruey-Shiong Tsay for their help in preparing this manuscript, and to Tom Leonard, a referee, and many other friends for helpful suggestions and comments.

REFERENCES

Abraham, B. and Box, G. E. P. (1978). Linear models and spurious observations. J. R. Statist. Soc. C, 27, 131-138.

Andrews, D. F. (1971a). A note on the selection of data transformations. Biometrika, 58, 249-254.

Andrews, D. F. (1971b). Significance tests based on residuals. Biometrika, 58, 139-148.

Anscombe, F. J. (1961). Examination of residuals. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 1-36. Berkeley and Los Angeles: University of California Press.

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. Technometrics, 5, 141-160.

Atkinson, A. C. (1973). Testing transformations to normality. J. R. Statist. Soc. B, 35, 473-479.

Bailey, S. P. and Box, G. E. P. (1980a). Modeling the nature and frequency of outliers. Technical Summary Report #2085, Mathematics Research Center, University of Wisconsin-Madison.

Bailey, S. P. and Box, G. E. P. (1980b). The duality of diagnostic checking and robustification in model building: Some considerations and examples. Technical Summary Report #2086, Mathematics Research Center, University of Wisconsin-Madison.

Barnard, G. (1978). Personal communication.

Box, G. E. P. (1976). Science and statistics. J. Amer. Statist. Ass., 71, 791-799.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Robustness in Statistics, 201-236, eds. Launer, R. L. and Wilkinson, G. N. New York, Academic Press.

Box, G. E. P. and Behnken, D. W. (1960). Some new three-level designs for the study of quantitative variables. Technometrics, 2, 455-475.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. J. R. Statist. Soc. B, 26, 211-243.

Box, G. E. P. and Draper, N. R. (1975). Robust designs. Biometrika, 62, 347-352.

Box, G. E. P. and Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

Box, G. E. P. and Newbold, P. (1971). Some comments on a paper by Coen, Gomme and Kendall. J. R. Statist. Soc. A, 134, 229-240.

Box, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes' theorem. Biometrika, 49, 419-432.

Box, G. E. P. and Tiao, G. C. (1964). A Bayesian approach to the importance of assumptions applied to the comparison of variances. Biometrika, 51, 153-167.

Box, G. E. P. and Tiao, G. C. (1968a). Bayesian analysis of means for the random effect model. J. Amer. Statist. Ass., 63, 174-181.

Box, G. E. P. and Tiao, G. C. (1968b). A Bayesian approach to some outlier problems. Biometrika, 55, 119-129.

Box, G. E. P. and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. Reading, Mass. Addison-Wesley.

Box, G. E. P. and Youle, P. V. (1955). The exploration and exploitation of response surfaces: An example of the link between the fitted surface and the basic mechanism of the system. Biometrics, 11, 287-323.

Chen, G. G. and Box, G. E. P. (1979). Further study of robustification via a Bayesian approach. Technical Summary Report #1998, Mathematics Research Center, University of Wisconsin-Madison.

Coen, P. J., Gomme, E. E. and Kendall, M. G. (1969). Lagged relationships in economic forecasting. J. R. Statist. Soc. A, 132, 133-152.

Cox, D. R. (1977). The Role of Significance Tests. Scand. J. Statist., 4, 49-70.

Dempster, A. P. (1971). Model searching and estimation in the logic of inference (with discussion). In Foundations of Statistical Inference, 56-81. Toronto, Holt, Rinehart and Winston.

Dempster, A. P. (1975). A subjectivist look at robustness. I.S.I. Bulletin, 46, 349-374.

Dixon, W. J. (1953). Processing data for outliers. Biometrics, 9, 74-89.

Draper, N. R. and Van Nostrand, R. C. (1977). Ridge regression: is it worthwhile? Technical Report 501, Department of Statistics, University of Wisconsin-Madison.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least square regression I. Biometrika, 37, 409-428.

Feller, W. (1968). An Introduction to Probability Theory and its Applications. Vol. 1, Wiley.

Fisher, R. A. (1935). The design of experiments, Oliver and Boyd, Edinburgh.

Geisser, S. (1971). The inferential use of predictive distributions (with discussion). In Foundations of Statistical Inference, 458-469. Toronto, Holt, Rinehart and Winston.

Geisser, S. (1975). The predictive sample reuse method with applications. J. Amer. Statist. Ass., 70, 320-328.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. J. Amer. Statist. Ass., 74, 153-160.

Good, I. J. (1956). The surprise index for the multivariate normal distribution. Ann. Math. Statist., 27, 1130-1135.

Gorman, J. W. and Toman, R. J. (1966). Selection of variables for fitting equations to data. Technometrics, 8, 27-51.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. J. R. Statist. Soc. B, 29, 83-100.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to non-orthogonal problems. Technometrics, 12, 69-82.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulations. Comm. in Statist., 4, 105-124.

Huber, P. J. (1977). Robust statistical procedures. Society for Industrial and Applied Mathematics, 27, Philadelphia.

Jeffreys, H. (1932). An alternative to the rejection of observations. Proc. Royal Soc., A, CXXXVII, 78-87.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1979). Interactive elicitation of opinion for a normal linear model. Technical Report 150, Dept. of Statistics, Carnegie-Mellon University.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes' Estimates for the Linear Model, (with discussion). J. R. Statist. Soc. B, 34, 1-41.

Pallesen, L. C. (1977). Studies in the analysis of serially dependent data. Ph.D. thesis, Department of Statistics, University of Wisconsin-Madison.

Popper, K. R. (1959). The Logic of Scientific Discovery. New York: Harper and Row.

Roberts, H. V. (1965). Probabilistic prediction. J. Amer. Statist. Ass., 60, 50-62.

Snedecor, G. W. and Cochran, W. G. (1967). Statistical Methods. Ames: Iowa State University Press.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium, 1, 197-206. Berkeley and Los Angeles: University of California Press.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniel and the history of robust estimation 1885-1920. J. Amer. Statist. Ass., 68, 872-879.

Theil, H. (1963). On the use of incomplete prior information in regression analysis. J. Amer. Statist. Ass., 58, 401-414.

Tiao, G. C. and Ali, M. M. (1971). Analysis of correlated random effects: linear model with two random components, Biometrika, 58, 37-51.

Tiao, G. C. and Box, G. E. P. (1980). An introduction to applied multiple time series analysis. Technical Report 582, Department of Statistics, University of Wisconsin-Madison.

Tukey, J. W. (1949). One degree of freedom for non-additivity, Biometrics, 5, 232-242.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 448-485. Stanford, Stanford University Press.

Zellner, A. and Tiao, G. C. (1964). Bayesian analysis of the regression model with autocorrelated errors. J. Amer. Statist. Ass., 59, 763-778.

GEPB/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2160	2. GOVT ACCESSION NO. AD-A096	3. RECIPIENT'S CATALOG NUMBER 645
4. TITLE (and Subtitle) SAMPLING AND BAYES' INFERENCE IN SCIENTIFIC MODELING AND ROBUSTNESS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) George E. P. Box		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE December 1980 13. NUMBER OF PAGES 43
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Technical Summary Rept.		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Iterative Learning; Model Building; Inference; Bayes' Theorem; Sampling Theory; Predictive Distribution; Diagnostic Checks; Transformations; Serial Correlation; Bad Values; Outliers; Robust Estimation.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Scientific learning is an iterative process employing Criticism and Estimation. Correspondingly the formulated model factors into two complimentary parts - a predictive part allowing model criticism, and a Bayes posterior part allowing estimation. Implications for significance tests, the theory of precise measurement, and for ridge estimates are considered. Predictive checking functions for transformation, serial correlation, bad values, and their relation with Bayesian options are considered. Robustness is seen from a Bayesian viewpoint and examples are given. For the bad value problem a comparison with M estimators is made.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 322200

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)